Generating quantum-measurement probabilities from an optimality principle

Johan A.K. Suykens

K.U. Leuven, ESAT-SCD/SISTA

Kasteelpark Arenberg 10

B-3001 Leuven (Heverlee), Belgium

Tel: 32/16/32 18 02 Fax: 32/16/32 19 70

Email: johan.suykens@esat.kuleuven.be

published: Phys. Rev. A 87, 052134 (2013)

http://link.aps.org/doi/10.1103/PhysRevA.87.052134

Abstract

An alternative formulation to the (generalized) Born rule is presented. It involves estimating an unknown model from a finite set of measurement operators on the state. An optimality principle is given that relates to achieving bounded solutions by regularizing the unknown parameters in the model. The objective function maximizes a lower bound on the quadratic Renyi classical entropy. The unknowns of the model in the primal are interpreted as transition witnesses. An interpretation of the Born rule in terms of fidelity is given with respect to transition witnesses for the pure state and POVM case. The models for generating quantum measurement probabilities apply to orthogonal projective measurements and POVM measurements, and to isolated and open systems with Kraus maps. A straightforward and constructive method is proposed for deriving the probability rule, which is based on Lagrange duality. An analogy is made with a kernel-based method for probability mass function estimation, for which similarities and differences are discussed. These combined insights from quantum mechanics, statistical modelling and machine learning provide an alternative way of generating quantum measurement probabilities.

1 Introduction

Quantum measurement is described by a set of measurement operators $\{M_i\}$ where the index i refers to possible measurement outcomes that may occur in the experiment [1, 2, 3, 4]. Following e.g. Postulate 3 in Nielsen & Chuang [1], for an isolated system the quantum measurement postulate states that if before measurement the state of a closed quantum system is $|\psi\rangle$, then the probability that the result i occurs is $p(i) = \left\langle \psi \middle| M_i^\dagger M_i \middle| \psi \right\rangle$, and the state after measurement is $M_i |\psi\rangle / \sqrt{\left\langle \psi \middle| M_i^\dagger M_i \middle| \psi \right\rangle}$ where the completeness relation $\sum_i M_i^\dagger M_i = I$ holds. Or alternatively for a system described by a density operator ρ the probability equals $p(i) = \mathrm{Tr}(M_i^\dagger M_i \rho)$ with state after measurement $\frac{M_i \rho M_i^\dagger}{\mathrm{Tr}(M_i^\dagger M_i \rho)}$. This postulate (i.e. the (generalized) Born rule [5]) has been a very successful and accurate recipe in its application. The von Neumann projected value measure [6] has been extended to a positive operator valued measure (POVM) [7, 8, 9, 10] and extensions to open systems have been made with Kraus maps [11]. However, from the viewpoint of why the probability rule takes this particular form or from where this expression could originate, it remains a topic of debate. Furthermore, with respect to quantum measurement several different interpretations exist, including the Copenhagen interpretation, many-worlds interpretation and several others [12, 13, 14].

A fundamental result is Gleason's theorem which states [15]: "*Let μ be a measure on the closed subspaces of a separable (real or complex) Hilbert space $\mathcal{H}$ of dimension at least three. There exists a positive semi-definite self-adjoint operator T of the trace class such that for all closed subspaces A of $\mathcal{H}$: $\mu(A) = \mathrm{Tr}(TP_A)$ where P_A is the orthogonal projection of $\mathcal{H}$ onto A*". According to Bell [16] the relevant corollary of Gleason's work is that the additivity requirement for expectation values of commuting operators cannot be met by dispersion free states. The proof of Gleason's theorem is however difficult to grasp and is only valid for projective measurements. A more elementary proof has been proposed in [17] and extensions have been made for POVM [18, 19]. Therefore Gleason's theorem and its extensions have contributed towards reducing the axiomatic basis of quantum mechanics. Another result in this direction but based on physical arguments has been made e.g. by Zurek [20].

The aim of this paper is to show that the (generalized) Born rule can be obtained as the optimal solution to an underlying model estimation problem, and therefore an alternative formulation can be given to the Born rule. A very common problem in model estimation is a classical regression problem where from given data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ with input data $x_i \in \mathbb{R}^d$ and output data $y_i \in \mathbb{R}$ one estimates a model $\hat{y} = f(x; w)$ with $\hat{y}$ the estimated output value and unknown parameters w. A simple case is a linear model $\hat{y} = w^T x + b$ where one estimates w, b from the given data $\mathcal{D}$. In this paper the quantum measurement problem is casted within such a simple setting. However, instead of having a supervised learning problem with given target values y_i as in a regression problem, in the quantum measurement problem before measurement, such values are unknown. We will therefore proceed by formulating an unsupervised learning problem (common examples of unsupervised learning studied in the fields of machine learning, neural networks and statistics are density estimation and clustering problems) with also y_i as unknowns to the problem.

The model specification is done then as follows. An unknown function is considered that maps $|M_i\psi\rangle$ to a possible outcome y_i for which it is imposed that $\sum_{i=1}^N y_i = 1$ with N the number of measurement outcomes. The unknown model is characterized then by an unknown bra $\langle w|$. In order to determine the unknown $\langle w|$ and y_i a constrained optimization problem is formulated. This enables to characterize the Lagrange dual problem, i.e. the problem in the Lagrange multipliers [21], which results into the Born rule. Currently, kernel methods for problems in supervised and unsupervised learning and beyond are often viewed in terms of primal and Lagrange dual model representations in methods of support vector machines [22, 24, 25, 23] and least squares support vector machines [26, 27]. A recent overview about the role of primal and Lagrange dual representations for a wide range of problems in supervised and unsupervised learning has been presented in [28].

In this model estimation approach to the Born rule the objective is to minimize $\langle w|w\rangle$. We will show how this leads to maximizing a lower bound on the quadratic Renyi classical entropy (collision entropy). In analogy with entanglement witnesses [29, 30], it is proposed to view $|w\rangle$ as a *transition witness*. At optimality it senses $|M_i\psi\rangle$ for $i = 1, ..., N$ and witnesses for the transition probabilities, along the physical interpretation of fidelity given by Jozsa in [31].

In this paper we will derive the Born rule in view of primal and dual model representations. Models are presented for the pure state and mixed state case. Both isolated and open systems are addressed. An analogy with a kernel-based model formulation for probability mass function (pmf) estimation is also presented because of the striking similarity with quantum measurement. The proofs of Theorem 2 and Lemma 1 are shown in appendix.

2 An underlying model and optimality principle to quantum measurement: the pure state case

Given an isolated quantum mechanical system with state vector $|\psi\rangle \in \mathbb{C}^d$ and measurement operators M_i for $i = 1, ..., N$, we consider mapping $|M_i\psi\rangle$ to an unknown value y_i through a function $f : \mathbb{C}^d \to \mathbb{C}$ by

$$y_i = \mathrm{Re}\left(f(|M_i\psi\rangle)\right) \tag{1}$$

where y_i is a real-valued outcome, for which the normalization $\sum_{i=1}^N y_i = 1$ is imposed with N the number of measurement outcomes. Throughout this paper we assume that N is finite. In this section we first consider measurement operators at a general level, including also non-Hermitian operators, and then specialize the results to orthogonal projective measurements. The unknown function f is represented in terms of an unknown bra $\langle w|$ by taking the following model:

$$y_i = \mathrm{Re}(\langle w|M_i\psi\rangle) \tag{2}$$

where in general the bra-ket $\langle w|M_i\psi\rangle$ can be complex valued and the set $\{|M_i\psi\rangle\}_{i=1}^N$ is given. Note that one also has $y_i = \mathrm{Re}(\langle w|M_i\psi\rangle) = \mathrm{Re}(\mathrm{Tr}(M_i\,|\psi\rangle\,\langle w|))$ using the cyclic trace property.

The unknown $|w\rangle$ in this model and the unknown y_i values are then determined after specifying the following constrained optimization problem (primal problem (P)):

$$\begin{array}{ll} \min\limits_{|w\rangle, y_i} & \frac{1}{2}\langle w|w\rangle \\ \text{subject to} & y_i = \mathrm{Re}(\langle w|M_i\psi\rangle), \qquad i = 1, ..., N \\ & \sum_{i=1}^N y_i = 1 \end{array} \tag{3}$$

where the outcomes y_i are normalized in the last equation. Note that at this point we do not impose the condition $y_i \geq 0$. This will be achieved through the assumption made on M_i.

Before presenting the results for orthogonal projective measurements, and POVM in the next section, we first state the following mathematical result for the model estimation problem (3).

Theorem 1. The optimal solution to the model estimation problem (3) is given by

$$y_i = \frac{\left\langle \psi \middle| \sum_j M_j^\dagger M_i \middle| \psi \right\rangle}{\left\langle \psi \middle| \sum_l \sum_j M_j^\dagger M_l \middle| \psi \right\rangle} \tag{4}$$

for $i = 1, ..., N$. For a probability rule interpretation $p(i) = y_i$ the assumption $\sum_j M_j^\dagger M_i \geq 0$ for $i = 1, ..., N$ is required.

Proof: In order to characterize the optimal solution to (3) we first rewrite the problem in terms of real-valued unknowns. The unknown $|w\rangle$ is expressed as $|w\rangle = \mathrm{Re}(|w\rangle) + \mathsf{i}\,\mathrm{Im}(|w\rangle)$ with real and imaginary parts $\mathrm{Re}(|w\rangle), \mathrm{Im}(|w\rangle) \in \mathbb{R}^d$. One has

$$\begin{aligned} \langle w|w\rangle &= \langle \mathrm{Re}(|w\rangle), \mathrm{Re}(|w\rangle)\rangle + \langle \mathrm{Im}(|w\rangle), \mathrm{Im}(|w\rangle)\rangle \\ \mathrm{Re}(\langle w|v\rangle) &= \langle \mathrm{Re}(|w\rangle), \mathrm{Re}(|v\rangle)\rangle + \langle \mathrm{Im}(|w\rangle), \mathrm{Im}(|v\rangle)\rangle \end{aligned}$$

where e.g. $|v\rangle = |M_i\psi\rangle$ can be taken. These expressions are made in terms of inner products that involve the real and imaginary parts of $|w\rangle$, $|v\rangle$.

The problem is then restated as follows with $\mathrm{Re}(|w\rangle), \mathrm{Im}(|w\rangle$ as independent unknowns:

$$\begin{aligned} \min_{\mathrm{Re}(|w\rangle),\mathrm{Im}(|w\rangle),y_i} \quad & \tfrac{1}{2}\langle \mathrm{Re}(|w\rangle), \mathrm{Re}(|w\rangle)\rangle + \tfrac{1}{2}\langle \mathrm{Im}(|w\rangle), \mathrm{Im}(|w\rangle)\rangle \\ \text{subject to} \quad & y_i = \langle \mathrm{Re}(|w\rangle), \mathrm{Re}(|M_i\psi\rangle)\rangle + \langle \mathrm{Im}(|w\rangle), \mathrm{Im}(|M_i\psi\rangle)\rangle\,, \qquad i = 1, ..., N \\ & \textstyle\sum_{i=1}^N y_i = 1. \end{aligned} \tag{5}$$

In order to characterize the solution the Lagrangian is constructed

$$\begin{aligned} &\mathcal{L}(\mathrm{Re}(|w\rangle), \mathrm{Im}(|w\rangle), y_i, \alpha_i, \beta) = \tfrac{1}{2}\langle \mathrm{Re}(|w\rangle), \mathrm{Re}(|w\rangle)\rangle + \tfrac{1}{2}\langle \mathrm{Im}(|w\rangle), \mathrm{Im}(|w\rangle)\rangle + \\ &\textstyle\sum_{i=1}^N \alpha_i (y_i - \langle \mathrm{Re}(|w\rangle), \mathrm{Re}(|M_i\psi\rangle)\rangle - \langle \mathrm{Im}(|w\rangle), \mathrm{Im}(|M_i\psi\rangle)\rangle) + \beta(\textstyle\sum_{i=1}^N y_i - 1) \end{aligned}$$

where $\alpha_i, \beta \in \mathbb{R}$ are Lagrange multipliers. Taking the conditions for optimality gives

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \operatorname{Re}(|w\rangle)} = 0 & \Rightarrow \quad \operatorname{Re}(|w\rangle) = \sum_{i=1}^N \alpha_i \operatorname{Re}(|M_i\psi\rangle) \\ \dfrac{\partial \mathcal{L}}{\partial \operatorname{Im}(|w\rangle)} = 0 & \Rightarrow \quad \operatorname{Im}(|w\rangle) = \sum_{i=1}^N \alpha_i \operatorname{Im}(|M_i\psi\rangle) \\ \dfrac{\partial \mathcal{L}}{\partial y_i} = 0 & \Rightarrow \quad \alpha_i + \beta = 0, i = 1, ..., N \\ \dfrac{\partial \mathcal{L}}{\partial \alpha_i} = 0 & \Rightarrow \quad y_i = \langle \operatorname{Re}(|w\rangle), \operatorname{Re}(|M_i\psi\rangle)\rangle + \langle \operatorname{Im}(|w\rangle), \operatorname{Im}(|M_i\psi\rangle)\rangle \, , i = 1, ..., N \\ \dfrac{\partial \mathcal{L}}{\partial \beta} = 0 & \Rightarrow \quad \sum_{i=1}^N y_i = 1. \end{cases}$$

The first two conditions can be combined into

$$|w\rangle = \operatorname{Re}(|w\rangle) + \mathsf{i} \operatorname{Im}(|w\rangle) = \sum_{i=1}^N \alpha_i \, |M_i\psi\rangle \, .$$

The fourth condition corresponds to $y_i = \operatorname{Re}(\langle w|M_i\psi\rangle) = \operatorname{Re}(\sum_{j=1}^N \alpha_j \left\langle \psi \middle| M_j^\dagger M_i \middle| \psi \right\rangle)$. Inserting the first four conditions into the last condition gives $\beta \sum_i \sum_j \left\langle \psi \middle| M_j^\dagger M_i \middle| \psi \right\rangle + 1 = 0$. Using this expression of β in the fourth condition gives (4).

The assumption $\sum_j M_j^\dagger M_i \geq 0$ for $i = 1, ..., N$ guarantees that $p(i) = y_i \geq 0$.

□

Remark 1. Note that one can also eliminate y_i from (3) which gives then the problem statement $\min_{|w\rangle} \frac{1}{2} \langle w|w\rangle$ s.t. $\sum_{i=1}^N \operatorname{Re}(\langle w|M_i\psi\rangle) = 1$. However, for interpreting the resulting model representation, (3) is more revealing, also in connection to kernel methods. The formulation as a constrained optimization problem (3) enables to express the underlying model (2) in terms of the Lagrange multipliers α_i (dual variables). The model $\mathcal{M}$ in (2) has then the following primal (P) and Lagrange dual (D) representation:

$$\mathcal{M} \begin{cases} \nearrow \quad (P): \quad y_i = \operatorname{Re}(\langle w|M_i\psi\rangle) \\ \searrow \quad (D): \quad y_i = \operatorname{Re}(\sum_{j=1}^N \alpha_j \left\langle \psi \middle| M_j^\dagger M_i \middle| \psi \right\rangle) = \frac{\langle \psi | \sum_j M_j^\dagger M_i | \psi \rangle}{\langle \psi | \sum_l \sum_j M_j^\dagger M_l | \psi \rangle}. \end{cases} \tag{6}$$

From the conditions for optimality it followed that $\alpha_i = -\beta$. Therefore at this level the expansion in the dual model representation is non-sparse.

Now, we connect Theorem 1 to the special case of orthogonal projective measurements. The following result directly follows from the expression $\sum_j M_j^\dagger M_i$ in Theorem 1. It derives the Born rule as a special case of Theorem 1.

Corollary 1 [orthogonal projective measurement]. For an observable $A = \sum_i a_i M_i$ where $M_j M_i = \delta_{ji} M_j$, $M_i^\dagger = M_i$ (orthogonal projective measurement) with completeness condition $\sum_j M_j = I$ and $\langle \psi | \psi \rangle = 1$, equation (4) simplifies to the Born rule

$$p(i) = y_i = \langle \psi | M_i | \psi \rangle \, . \tag{7}$$

In this case at optimality one has $|w\rangle = |\psi\rangle$.

Remark 2. The above results apply to a pure state $|\psi\rangle$. For mixed states $\rho = \sum_i p_i \, |\psi_i\rangle \, \langle \psi_i|$ one has the probability rule $p(m) = \sum_i p(m|i) p_i = \mathrm{Tr}(M_m \rho)$ with $p(m|i) = \langle \psi_i | M_m | \psi_i \rangle$ [1].

Remark 3. The use of a regularization term like $\langle w|w\rangle$ in (3) is very common in many methods of function estimation, including splines, support vector machines and kernel methods, ridge regression, weight decay in neural networks and others [33, 34, 35], though for a real-valued vector w. Regularization terms are important then as a mechanism to control the model complexity and achieving a good generalization and predictive performance. More specifically in the context of function estimation in a Hilbert space, one often considers the function estimation in a reproducing kernel Hilbert space (RKHS) with a regularization term, resulting into bounded evaluation functionals [32].

Let us now look further into the role of $\langle w|w\rangle = \mathrm{Tr}(|w\rangle \, \langle w|)$ in the objective. We discuss its role in view of (i) normalization and probability interpretation; (ii) characterization of a lower bound on collision classical entropy; (iii) a new notion of transition witness; (iv) the physical interpretation of fidelity.

The role of $\langle w|w\rangle$ in (3) can be further understood from the application of Cauchy-Schwarz:

$$| \, \langle w | M_i \psi \rangle \, |^2 \leq \langle w|w\rangle \left\langle \psi \middle| M_i^\dagger M_i \middle| \psi \right\rangle . \tag{8}$$

For given $|\psi\rangle$ and M_i, the minimization of $\langle w|w\rangle$ leads to bounding $| \, \langle w|M_i\psi\rangle \, |$. For $\langle \psi|\psi\rangle = 1$, at optimality $|w\rangle = -\beta \sum_{i=1}^N |M_i \psi\rangle)$ holds, which results into $\langle \psi|M_i|\psi\rangle \leq 1$. In this way the minimization of $\langle w|w\rangle$ leads to a normalization and probability interpretation.

The minimization of $\langle w|w\rangle$ leads to maximizing a lower bound on the quadratic Renyi classical entropy $H_2 = -\log \sum_i p(i)^2$ which is a lower bound on the Shannon classical entropy $H_1 = -\sum_i p(i) \log p(i)$. From (8) and $p(i)^2 = |\mathrm{Re}(\langle w|M_i\psi\rangle)|^2 \leq |\mathrm{Re}(\langle w|M_i\psi\rangle)|^2 + |\mathrm{Im}(\langle w|M_i\psi\rangle)|^2$ we obtain

$$H_2 \geq -\log \langle w|w\rangle \tag{9}$$

due to the fact that $\sum_i \left\langle \psi \middle| M_i^\dagger M_i \middle| \psi \right\rangle = 1$. At optimality this yields $H_2 \geq 0$ from $\langle w|w\rangle = 1$.

In analogy with the notion of entanglement witnesses [29, 30] (where for any entangled state ρ, there exists a Hermitian operator A such that $\mathrm{Tr}(A\,\rho) < 0$, while $\mathrm{Tr}(A\,\sigma) \geq 0$ for all separable states σ), it is proposed here to call $|w\rangle$ a transition witness.

At optimality $|w\rangle$ is in the superposition $|w\rangle = -\beta \sum_{i=1}^N |M_i\psi\rangle$ which senses the correlations with $|M_i\psi\rangle$. The completeness condition $\sum_i M_i = I$ results into $|w\rangle = |\psi\rangle$. Hence the state itself serves as a self-reference and witness for the transition probability. Considering the computational basis $M_i = |i\rangle\langle i|$, at optimality one obtains

$$p(i) = \langle\psi|i\rangle\langle\psi|i\rangle = |\langle\psi|i\rangle|^2 = F(|\psi\rangle\langle\psi|, |i\rangle\langle i|) = F(|w\rangle\langle w|, |i\rangle\langle i|) \tag{10}$$

where F denotes the fidelity [31, 1]. $F(|\psi_1\rangle\langle\psi_1|, \rho_2) = \langle\psi_1|\rho_2|\psi_1\rangle$ is the fidelity which measures the closeness between a pure state $|\psi_1\rangle$ and a density matrix ρ_2. The physical interpretation of (10) based on [31] is that it is the probability that $|i\rangle\langle i|$ passes the yes/no test of being pure state $|\psi\rangle$ with as test the measurement of observable $|\psi\rangle\langle\psi| = |w\rangle\langle w|$.

Example 1. Taking a qubit example $|\psi\rangle = a\,|0\rangle + b\,|1\rangle$ with $a, b \in \mathbb{C}$, $\langle\psi|\psi\rangle = 1$ and $M_i = |i\rangle\langle i|$, at optimality the transition witness takes the value $|w\rangle = -\beta \sum_{i=1}^N |M_i\psi\rangle = |\psi\rangle$ which is equal to the state $|\psi\rangle$ itself. At optimality one has $p(0) = \langle w|M_0\psi\rangle = \langle\psi|0\rangle\langle 0|\psi\rangle = |a|^2$ and $p(1) = \langle w|M_1\psi\rangle = \langle\psi|1\rangle\langle 1|\psi\rangle = |b|^2$ which are the probabilities that $|0\rangle\langle 0|$ and $|1\rangle\langle 1|$ pass the yes/no test of being pure state $|\psi\rangle$. At optimality one has that $\langle w|w\rangle = 1$.

3 A model related to mixed states

Let us now consider a model that operates on $\eta \in \mathbb{C}^{d\times d}$ which is connected to a given density matrix $\rho \in \mathbb{C}^{d\times d}$ by $\rho = \eta\eta^\dagger$. For this purpose, based on the eigenvalue decomposition

$\rho = USU^\dagger$, we define $\eta = US^{1/2}$.

Given operators $\{\Pi_i\}$, we consider now mapping $|\Pi_i\eta\rangle$ to an unknown value y_i through a function $g : \mathbb{C}^{d\times d} \to \mathbb{C}^{d\times d}$ by

$$y_i = \mathrm{Re}\,\mathrm{Tr}\, g(|\Pi_i\eta\rangle) \tag{11}$$

where y_i is a real-valued outcome, for which the normalization $\sum_{i=1}^N y_i = 1$ is imposed with N the number of measurement outcomes. The unknown function g is represented in terms of an unknown matrix $W \in \mathbb{C}^{d\times d}$ by taking the following model:

$$y_i = \mathrm{Re}\,\mathrm{Tr}\,(W^\dagger\Pi_i\eta) \tag{12}$$

where in general $\mathrm{Tr}(W^\dagger\Pi_i\eta)$ can be complex valued and the set $\{|\Pi_i\eta\rangle\}_{i=1}^N$ is given.

The following optimization problem is considered then

$$\begin{array}{ll} \min\limits_{W,y_i} & \frac{1}{2}\mathrm{Tr}(W^\dagger W) \\ \text{subject to} & y_i = \mathrm{Re}\,\mathrm{Tr}\,(W^\dagger\Pi_i\eta), \qquad i = 1,...,N \\ & \sum_{i=1}^N y_i = 1. \end{array} \tag{13}$$

Note that one has the following properties for the Hilbert-Schmidt inner product [36, 1]

$$\mathrm{Tr}(A^\dagger B) \;=\; \textstyle\sum_j \langle a_j|b_j\rangle = \sum_j (Ae_j)^\dagger Be_j = \sum_j \mathrm{Tr}(A^\dagger B\,|e_j\rangle\,\langle e_j|)$$

with $e_j = [0...1...0]^T$ with value 1 at position j and $A = [|a_1\rangle\,...\,|a_d\rangle]$, $B = [|b_1\rangle\,...\,|b_d\rangle]$.

Therefore the problem (13) can be stated also as

$$\begin{array}{ll} \min\limits_{|w_j\rangle,y_i} & \frac{1}{2}\sum_j \langle w_j|w_j\rangle \\ \text{subject to} & y_i = \mathrm{Re}(\sum_j \langle w_j|\Pi_i\eta e_j\rangle), \qquad i = 1,...,N \\ & \sum_{i=1}^N y_i = 1 \end{array} \tag{14}$$

which establishes the connection with the previous model estimation (3). The following general mathematical result is then obtained.

Theorem 2. The optimal solution to the model estimation problem (14) is given by the values

$$y_i = \frac{\mathrm{Tr}(\sum_l \Pi_l^\dagger\Pi_i\rho)}{\mathrm{Tr}(\sum_k\sum_l \Pi_l^\dagger\Pi_k\rho)} \tag{15}$$

for $i = 1, ..., N$. For a probability rule interpretation $p(i) = y_i$ the assumption $\sum_l \Pi_l^\dagger \Pi_i \geq 0$ for $i = 1, ..., N$ is required.

For POVMs the following results immediately as a special case of Theorem 2:

Corollary 2 [POVM]. For a POVM (positive-operator valued measure) decomposition with $\sum_l \Pi_l = I$ and $\Pi_i \geq 0$ for $i = 1, ..., N$ the probability rule (15) reduces to

$$p(i) = \mathrm{Tr}(\Pi_i \rho) \tag{16}$$

for $i = 1, ..., N$.

Corollary 3. Applying the model (13) to $\{\Pi_i \rho\}_{i=1}^N$ instead of to $\{\Pi_i \eta\}_{i=1}^N$ would result for a POVM into $y_i = \frac{\mathrm{Tr}(\Pi_i \rho^2)}{\mathrm{Tr}(\rho^2)}$. Only for pure states, for which $\rho^2 = \rho$ holds in that case, this reduces then to $p(i) = \mathrm{Tr}(\Pi_i \rho)$.

Remark 4. The result is also applicable to open systems described by Kraus maps $\mathcal{E}(\rho) = \sum_i E_i \rho E_i^\dagger$ with trace-preserving property $\sum_i E_i^\dagger E_i = I$ with a system coupled to the environment being in a product state $\rho \otimes \rho_{\mathrm{env}}$. One has then $p(i) = \mathrm{Tr}(E_i^\dagger E_i \rho)$ where $E_i^\dagger E_i$ plays the role of Π_i in (16). One has $\mathcal{E}(\rho) = \sum_i p(i) \rho_i$ with $\rho_i = \frac{E_i \rho E_i^\dagger}{\mathrm{Tr}(E_i \rho E_i^\dagger)}$ [1]. More generally one considers a non-trace-preserving property $\sum_i E_i^\dagger E_i \leq I$ [1]. Recent work on taking a measurement apparatus into account is e.g. [37]. Information gain and balance have been discussed in [38, 39, 40].

In order to understand the role of the minimization $\mathrm{Tr}(W^\dagger W)$ in the objective in (13), again one can apply Cauchy-Schwarz. It gives

$$|\mathrm{Tr}(W^\dagger \Pi_i \eta)|^2 \leq \mathrm{Tr}(W^\dagger W)\, \mathrm{Tr}(\eta^\dagger \Pi_i^\dagger \Pi_i \eta). \tag{17}$$

At optimality one has $|w_j\rangle = -\beta \sum_{i=1}^N |\Pi_i \eta e_j\rangle$ which gives then $|\mathrm{Tr}(\Pi_i \rho)|^2 \leq \mathrm{Tr}(\Pi_i^\dagger \Pi_i \rho)$. In the case of orthogonal projective measurements, Cauchy-Schwarz gives then at optimality the probability interpretation $\mathrm{Tr}(\Pi_i \rho) \leq 1$.

In this case the minimization of $\mathrm{Tr}(W^\dagger W)$ in (13) is maximizing a lower bound on the collision classical entropy H_2. The following result directly follows from (17).

Corollary 4 [Collision classical entropy]. For the model (13) the following holds

$$H_2 \geq -\log \mathrm{Tr}(W^\dagger W) - \log \mathrm{Tr}(\sum_i \Pi_i^\dagger \Pi_i \rho) \tag{18}$$

At optimality one obtains for POVM that $W = \eta$ such that $\log \mathrm{Tr}(W^\dagger W) = 0$. In that case $H_2 \geq -\log \mathrm{Tr}(\sum_i \Pi_i^\dagger \Pi_i \rho)$ holds.

In the mixed state case, the transition witnesses $|w_j\rangle$ for $j = 1, ..., d$ take at optimality the superposition form $|w_j\rangle = -\beta \sum_{i=1}^N |\Pi_i \eta e_j\rangle$ and sense the correlations with $|\Pi_i \eta e_j\rangle$ for $i = 1, ..., N$. At optimality, these are aggregated by summing them up to $p(i)$.

Corollary 5 [POVM Fidelity]. For the model (13) and POVM one has at optimality $W = \eta$. As a result the generalized Born rule can be expressed in terms of fidelity with respect to the transition witnesses $|w_j\rangle$:

$$p(i) = \sum_j \langle w_j | \Pi_i | w_j \rangle = \sum_j F(|w_j\rangle \langle w_j|, \Pi_i). \tag{19}$$

The latter expression (19) can be expressed in terms of a $N \times d$ matrix of fidelities:

$$\begin{bmatrix} p(1) \\ \vdots \\ p(N) \end{bmatrix} = \begin{bmatrix} F(|w_1\rangle \langle w_1|, \Pi_1) & ... & F(|w_d\rangle \langle w_d|, \Pi_1) \\ \vdots & & \vdots \\ F(|w_1\rangle \langle w_1|, \Pi_N) & ... & F(|w_d\rangle \langle w_d|, \Pi_N) \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}. \tag{20}$$

The physical interpretation is that Π_i passes the yes/no tests of being pure state $|w_j\rangle$ with as tests the measurements of observables $|w_j\rangle \langle w_j|$ for $j = 1, ..., d$. In [41] any operator η that satisfies $\rho = \eta\eta^\dagger$ has been called amplitude of ρ. It has been shown [41] that the change $\eta \to \eta U$ with U unitary is a gauge transformation with respect to a natural gauge potential related to the Bures Riemann metric. Because $W = \eta$ holds in the POVM case this property also applies here to the transition witnesses $|w_j\rangle$ that constitute W.

4 Analogy with a kernel-based probability mass function estimation method

We show now an analogy with a related but different problem of probability mass function estimation on given real-valued data $\mathcal{X} = \{x_i\}_{i=1}^N$ with $x_i \in \mathbb{R}^d$. We consider the following primal problem

$$\begin{array}{ll} \min\limits_{w, y_i} & \frac{1}{2} \langle w, w \rangle \\ \text{subject to} & y_i = \langle w, \varphi(x_i) \rangle \qquad i = 1, ..., N \\ & \sum_{i=1}^N y_i = 1 \end{array} \tag{21}$$

with $w \in \mathbb{R}^h$ the unknown of the model and feature map $\varphi : \mathbb{R}^d \to \mathbb{R}^h$ where h is the dimension of the high dimensional feature space. Such constrained optimization problem formulations are common for methods as support vector machines and least squares support vector machines [22, 28] in statistical estimation and machine learning problems of supervised and unsupervised learning. The formulation here relates to a probability mass function estimation problem [43]. The formulation consists of equality constraints for the underlying model as in least squares support vector machines, while in support vector machines inequality constraints are used commonly (a different primal-dual characterization for a class of Parzen estimators for a different problem of probability density function estimation [42] has been given in [44]).

Lemma 1. The optimal solution to the probability mass function estimation problem (21) is given by

$$y_i = \frac{\sum_{j=1}^N K(x_j, x_i)}{\sum_{i=1}^N \sum_{j=1}^N K(x_j, x_i)} \tag{22}$$

for $i = 1, ..., N$ where a positive definite kernel $K(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$ $(\forall x_i, x_j \in \mathcal{X})$ is employed.

Remark 5. Taking a positive definite kernel K guarantees the existence of a feature map φ, according to the Mercer theorem [45]. A possible choice for K to be used in (22) is a Gaussian kernel $K(x, z) = \exp(-\frac{\|x-z\|_2^2}{\sigma^2})$ such that one has $0 \leq y_i \leq 1$, $i = 1, ..., N$ and $\sum_{i=1}^N y_i = 1$.

Remark 6. The analogy between quantum measurement and this specific kernel-based pmf estimation method is further summarized in Table 1. It shows similarities and differences between the two problems. While in the kernel-based pmf estimation method the use of a positive definite kernel $K(x_i, x_j) = \langle \varphi(x_i), \varphi(x_j) \rangle$ enables to obtain the kernel-based representation in the dual, in quantum measurement the property $\langle i|j \rangle = \delta_{ij}$ (for the choice $M_i = |i\rangle \langle i|$) plays a key role in obtaining the simpler expression $y_i = \langle \psi | M_i | \psi \rangle$ in the dual representation.

	Quantum measurement pure state	Quantum measurement POVM	pmf estimation kernel-based method
Given	$\{\|M_i\psi\rangle\}_{i=1}^N$, $\|\psi\rangle \in \mathbb{C}^d$	$\{\|\Pi_i\eta\rangle\}_{i=1}^N$, $\eta \in \mathbb{C}^{d\times d}, \rho = \eta\eta^\dagger$	$\{x_i\}_{i=1}^N$, $x_i \in \mathbb{R}^d$
Model (primal)	$y_i = \mathrm{Re}(\langle w\|M_i\psi\rangle)$	$y_i = \mathrm{Re}\,\mathrm{Tr}\,(W^\dagger\Pi_i\eta)$	$y_i = \langle w, \varphi(x_i)\rangle$
	pure state $\|\psi\rangle$	η from mixed state ρ	feature map φ
Model (unknown)	$w \in \mathbb{C}^d$	$W \in \mathbb{C}^{d\times d}$	$w \in \mathbb{R}^h$
Transition witness	$\|w\rangle = -\beta\sum_i \|M_i\psi\rangle$	$\|w_j\rangle = -\beta\sum_i \|\Pi_i\eta e_j\rangle$	$w = -\beta\sum_i \varphi(x_i)$
Fidelity	$y_i = F(\|w\rangle\langle w\|, \|i\rangle\langle i\|)$	$y_i = \sum_j F(\|w_j\rangle\langle w_j\|, \Pi_i)$	-
Collision entropy	$H_2 \geq 0$	$H_2 \geq -\log\mathrm{Tr}(\sum_i \Pi_i^\dagger\Pi_i\rho)$	$H_2 \geq \log\sum_{ij} K(x_i,x_j) - \log\sum_i K(x_i,x_i$
Normalization	$\sum_{i=1}^N y_i = 1$	$\sum_{i=1}^N y_i = 1$	$\sum_{i=1}^N y_i = 1$
Model (dual)	$y_i = \langle\psi\|M_i\|\psi\rangle$	$y_i = \mathrm{Tr}(\Pi_i\rho)$	$y_i = \frac{\sum_{j=1}^N K(x_j,x_i)}{\sum_{i=1}^N\sum_{j=1}^N K(x_j,x_i)}$
Key property	$\langle i\|j\rangle = \delta_{ij}$ (if $M_i = \|i\rangle\langle i\|$)	$\sum_i \Pi_i = I$, $\Pi_i \geq 0$	$K(x_i,x_j) = \langle\varphi(x_i),\varphi(x_j)\rangle$

Table 1: An analogy between quantum measurement and a specific kernel-based method for probability mass function estimation, both with primal and Lagrange dual model representations.

5 Conclusions

An alternative formulation to the Born rule has been presented in this paper through an optimization principle, which provides an easy to grasp interpretation. Unlike Gleason's theorem it yields the Born rule also for dimensions lower than three. The dimensionality of the space has been assumed to be finite dimensional with a finite number of measurement operators. The results apply to isolated and open systems. It is a challenge to further extend the results to the infinite dimensional case.

While optimality principles and Lagrangians are abundantly present in physics, the Born rule has not been characterized in such a way. For example, in classical mechanics for Newton's second law of motion one *derives* the equation of motion as a stationary solution from a variational principle (the least action principle). In a similar spirit, in this paper

we started from the fact that nature acts according to the (generalized) Born rule $p(i) = \mathrm{Tr}(\Pi_i \rho)$ and we wanted then to *derive* this from an optimality principle. The interpretation of the primal problem has been related to the existence of transition witnesses, in analogy with the expression of entanglement witnesses. A characterization of a lower bound on the collision classical entropy has been given. The minimization of the objective in the optimization problem results into maximizing this lower bound. At optimality the dual problem yields the (generalized) Born rule. The Born rule has been also expressed in terms of fidelity with respect to transition witnesses, both for the pure state and POVM case.

Convex optimization has already been playing a role in e.g. designing optimal quantum detectors [46], quantum tomography via compressed sensing [47] and entanglement witnesses [29, 30]. We have shown that such an approach also enables to provide an alternative way for generating quantum measurement probabilities, both for isolated and open quantum systems.

Acknowledgments. The author thanks the reviewers for constructive comments and acknowledges support from KU Leuven, the Flemish government, FWO, the Belgian federal science policy office and the European Research Council (CoE EF/05/006, GOA MANET, IUAP DYSCO, FWO G.0377.12, ERC AdG A-DATADRIVE-B).

References

[1] M.A. Nielsen, I.L. Chuang, *Quantum Computation and Quantum Information*, (Cambridge University Press, Cambridge, 2000).

[2] J. Preskill, *Quantum Information and Computation*, (Lecture Notes for Physics 229, California Institute of Technology, 1998).

[3] A. Peres, *Quantum Theory: Concepts and Methods*, (Kluwer, Boston, 1995).

[4] C.W. Helstrom, *Quantum Detection and Estimation Theory*, (Academic Press, New York, 1976).

[5] M. Born, *Z. Phys.*, **37**(12), 863 (1926).

[6] J. von Neumann, *Mathematical Principles of Quantum Mechanics*, (Princeton University Press, 1955).

[7] E.B. Davies, J.T. Lewis, *Commun. Math. Phys.*, **17**, 239 (1970).

[8] E.B. Davies, *Quantum Theory of Open Systems*, (Academic Press, London, 1976).

[9] A.S. Holevo, *Probabilistic and Statistical Aspects of Quantum Theory*, (Series in Statistics and Probability, North-Holland, Amsterdam, 1982).

[10] A.S. Holevo, *Statistical Structure of Quantum Theory*, (Lect. Notes Phys. m67, Springer, Berlin 2001).

[11] K. Kraus, *States, Effects and Operations*, (Springer-Verlag, Berlin, 1983).

[12] A. Zeilinger, *Nature*, **408**, 639 (2000).

[13] M. Tegmark, J.A. Wheeler, *Scientific American*, **284**:68 (2001).

[14] J.A. Wheeler, W.H. Zurek, *Quantum Theory and Measurement*, (Princeton University Press, 1983)

[15] A.M. Gleason, *J. Math. Mech.*, **6**(6), 885 (1957).

[16] J.S. Bell, *Rev. Mod. Phys.*, **38**(3), 447 (1966).

[17] R. Cooke, M. Keane, W. Moran, *Math. Proc. Camb. Phil. Soc.*, **98**, 117 (1985).

[18] C.M. Caves, C.A. Fuchs, K. Manne, J.M. Renes, *Found. Phys.*, **34**, 193 (2004).

[19] P. Busch, *Phys. Rev. Lett.*, **91**(12), 120403 (2003).

[20] W.H. Zurek, *Phys. Rev. Lett.*, **90**(12), 120404 (2003).

[21] S. Boyd, L. Vandenberghe, *Convex Optimization*, (Cambridge University Press, 2004).

[22] C. Cortes, V. Vapnik, *Machine Learning*, **20**, 273 (1995).

[23] V. Vapnik, *Statistical Learning Theory*, (Wiley, New York, 1998).

[24] B. Schölkopf, A. Smola, *Learning with Kernels*, (MIT Press, Cambridge, MA, 2002).

[25] J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Analysis*, (Cambridge University Press, 2004).

[26] J.A.K. Suykens, J. Vandewalle, *Neural Processing Letters*, **9**(3), 293 (1999).

[27] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, *Least Squares Support Vector Machines*, (World Scientific, Singapore, 2002).

[28] J.A.K. Suykens, C. Alzate, K. Pelckmans, *Statistics Surveys,* **4**, 148 (2010).

[29] O. Guhne, G. Toth, *Phys. Rep.*, **474**, 1 (2009).

[30] A.C. Doherty, P.A. Parrilo, F.M. Spedalieri, *Phys. Rev. A*, **69**, 022308 (2004)

[31] R. Jozsa, *J. Mod. Opt.*, **41**, 2315 (1994).

[32] F. Cucker, S. Smale, *Bull. Amer. Math. Soc.*, **39**(1), 1 (2002).

[33] G. Wahba, *Spline Models for Observational Data*, (Series Appl. Math., 59, SIAM, 1990).

[34] T. Poggio, F. Girosi, *Proceedings of the IEEE*, **78**(9), 1481 (1990).

[35] D. MacKay, *Information Theory, Inference, and Learning Algorithms*, (Cambridge University Press, 2003).

[36] M. Keyl, *Phys. Rep.*, **369**, 431 (2002).

[37] T. Amri, J. Laurat, C. Fabre, *Phys. Rev. Lett.* **106**, 020502 (2011)

[38] M. Ozawa, *J. Math. Phys.* **27**, 759 (1986)

[39] F. Buscemi, M. Hayashi, M. Horodecki, *Phys. Rev. Lett.* **100**, 210504 (2008)

[40] S. Luo, *Phys. Rev. A*, **82**, 052103 (2010)

[41] A. Uhlmann, *Found Phys*, **41**, 288 (2011)

[42] A.J. Izenman, *J. Am. Stat. Assoc.*, **86**(413), 205 (1991).

[43] L. Wasserman, *All of Statistics*, (Springer, New York, 2004).

[44] K. Pelckmans, J.A.K. Suykens, B. De Moor, A Risk Minimization Principle for a Class of Parzen Estimators, in *Proc. of the Neural Information Processing Systems (NIPS 2007)*, Vancouver, Canada, Dec. 2007.

[45] J. Mercer, *Philos. Trans. Roy. Soc. London*, **209**, 415 (1909).

[46] Y.C. Eldar, A. Megretski, G.C. Verghese, *IEEE Trans. Inf. Th.*, **49**(4), 1007 (2003).

[47] S.T. Flammia, D. Gross, Y.-K. Liu, J. Eisert, *New J. Phys.* **14** 095022 (2012).

A Proof of Theorem 2 and Lemma 1

Proof of Theorem 2

As in the proof of Theorem 1, we first rewrite the problem in terms of real-valued unknowns. The unknown $|w_j\rangle$ are expressed as $|w_j\rangle = \mathrm{Re}(|w_j\rangle) + \mathsf{i}\,\mathrm{Im}(|w_j\rangle)$ with real and imaginary parts $\mathrm{Re}(|w_j\rangle), \mathrm{Im}(|w_j\rangle) \in \mathbb{R}^d$.

The problem is then restated as

$$\begin{array}{ll} \min\limits_{\mathrm{Re}(|w_j\rangle),\mathrm{Im}(|w_j\rangle),y_i} & \frac{1}{2}\sum_j \langle \mathrm{Re}(|w_j\rangle), \mathrm{Re}(|w_j\rangle)\rangle + \frac{1}{2}\sum_j \langle \mathrm{Im}(|w_j\rangle), \mathrm{Im}(|w_j\rangle)\rangle \\ \text{subject to} & y_i = \sum_j \langle \mathrm{Re}(|w_j\rangle), \mathrm{Re}(|\Pi_i \eta e_j\rangle)\rangle + \sum_j \langle \mathrm{Im}(|w_j\rangle), \mathrm{Im}(|\Pi_i \eta e_j\rangle)\rangle\,, \qquad i = 1, ..., N \\ & \sum_{i=1}^N y_i = 1. \end{array} \tag{23}$$

The corresponding Lagrangian is

$$\mathcal{L}(\mathrm{Re}(|w_j\rangle), \mathrm{Im}(|w_j\rangle), y_i, \alpha_i, \beta) = \frac{1}{2}\sum_j \langle \mathrm{Re}(|w_j\rangle), \mathrm{Re}(|w_j\rangle)\rangle + \frac{1}{2}\sum_j \langle \mathrm{Im}(|w_j\rangle), \mathrm{Im}(|w_j\rangle)\rangle + \sum_{i=1}^N \alpha_i (y_i - \sum_j \langle \mathrm{Re}(|w_j\rangle), \mathrm{Re}(|\Pi_i \eta e_j\rangle)\rangle - \sum_j \langle \mathrm{Im}(|w_j\rangle), \mathrm{Im}(|\Pi_i \eta e_j\rangle)\rangle) + \beta(\sum_{i=1}^N y_i - 1)$$

where $\alpha_i, \beta \in \mathbb{R}$ are Lagrange multipliers. Taking the conditions for optimality gives

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial\, \mathrm{Re}(|w_j\rangle)} = 0 \;\Rightarrow\; \mathrm{Re}(|w_j\rangle) = \sum_{i=1}^N \alpha_i \mathrm{Re}(|\Pi_i \eta e_j\rangle) \\ \dfrac{\partial \mathcal{L}}{\partial\, \mathrm{Im}(|w_j\rangle)} = 0 \;\Rightarrow\; \mathrm{Im}(|w_j\rangle) = \sum_{i=1}^N \alpha_i \mathrm{Im}(|\Pi_i \eta e_j\rangle) \\ \dfrac{\partial \mathcal{L}}{\partial y_i} = 0 \;\Rightarrow\; \alpha_i + \beta = 0, i = 1, ..., N \\ \dfrac{\partial \mathcal{L}}{\partial \alpha_i} = 0 \;\Rightarrow\; y_i = \sum_j \langle \mathrm{Re}(|w_j\rangle), \mathrm{Re}(|\Pi_i \eta e_j\rangle)\rangle + \sum_j \langle \mathrm{Im}(|w_j\rangle), \mathrm{Im}(|\Pi_i \eta e_j\rangle)\rangle\,, i = 1, ..., N \\ \dfrac{\partial \mathcal{L}}{\partial \beta} = 0 \;\Rightarrow\; \sum_{i=1}^N y_i = 1. \end{cases}$$

The first two conditions can be combined into $|w_j\rangle = \mathrm{Re}(|w_j\rangle) + \mathsf{i}\,\mathrm{Im}(|w_j\rangle) = \sum_{i=1}^N \alpha_i |\Pi_i \eta e_j\rangle = -\beta \sum_{i=1}^N |\Pi_i \eta e_j\rangle$. The expression for y_i becomes then

$$\begin{aligned} y_i &= \mathrm{Re}(\textstyle\sum_j \langle w_j | \Pi_i \eta e_j\rangle) \\ &= -\beta \mathrm{Re} \textstyle\sum_j \left\langle e_j \middle| \eta^\dagger \sum_l \Pi_l^\dagger \Pi_i \eta \middle| e_j \right\rangle \\ &= -\beta \mathrm{Tr}(\textstyle\sum_l \Pi_l^\dagger \Pi_i \eta \sum_j e_j e_j^\dagger \eta^\dagger) \\ &= -\beta \mathrm{Tr}(\textstyle\sum_l \Pi_l^\dagger \Pi_i \rho) \end{aligned}$$

after application of the cyclic trace property. From $\sum_{i=1}^{N} y_i = 1$ it follows that $\beta = -1/\mathrm{Tr}(\sum_k \sum_l \Pi_l^\dagger \Pi_k \rho)$ such that (15) is obtained. The assumption $\sum_l \Pi_l^\dagger \Pi_i \geq 0$ for $i = 1, ..., N$ guarantees that $p(i) = y_i \geq 0$.

Proof of Lemma 1

The Lagrangian for (21) is given by

$$\mathcal{L}(w, y_i, \alpha_i, \beta) = \frac{1}{2} \langle w, w \rangle + \sum_{i=1}^{N} \alpha_i (y_i - \langle w, \varphi(x_i) \rangle) + \beta (\sum_{i=1}^{N} y_i - 1)$$

with $\alpha_i, \beta \in \mathbb{R}$ Lagrange multipliers. The conditions for optimality are

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0 & \Rightarrow \quad w = \sum_{i=1}^{N} \alpha_i \varphi(x_i) \\ \frac{\partial \mathcal{L}}{\partial y_i} = 0 & \Rightarrow \quad \alpha_i + \beta = 0, i = 1, ..., N \\ \frac{\partial \mathcal{L}}{\partial \alpha_i} = 0 & \Rightarrow \quad y_i = \langle w, \varphi(x_i) \rangle, i = 1, ..., N \\ \frac{\partial \mathcal{L}}{\partial \beta} = 0 & \Rightarrow \quad \sum_{i=1}^{N} y_i = 1. \end{cases}$$

Elimination of w, α_i, β gives (22). It also follows that $\beta = -1/\sum_{ij} K(x_i, x_j)$.